IL NUOVO CIMENTO **40 C** (2017) 73
DOI 10.1393/ncc/i2017-17073-x

COLLOQUIA: IFAE 2016

Reconfigurable PCI Express cards for low-latency data transport in HEP experiments

R. AMMENDOLA(1), A. BIAGIONI(2), P. CRETARO(2), O. FREZZA(2), G. LAMANNA(3), F. LO CICERO(2), A. LONARDO(2), M. MARTINELLI(2), P. S. PAOLUCCI(2), E. PASTORELLI(2), L. PONTISSO(4), F. SIMULA(2) and P. VICINI(2)

(1) *INFN, Sezione di Roma Tor Vergata - Roma, Italy*
(2) *INFN, Sezione di Roma Sapienza - Roma, Italy*
(3) *INFN, Laboratori Nazionali di Frascati - Frascati (Roma), Italy*
(4) *INFN, Sezione di Pisa - Pisa, Italy*

received 17 October 2016

Summary. — State-of-the-art technology supports the High Energy Physics community in addressing the problem of managing an overwhelming amount of experimental data. From the point of view of communication between the detectors' readout system and computing nodes, the critical issues are the following: latency, moving data in a deterministic and low amount of time; bandwidth, guaranteeing the maximum capability of the link and communication protocol adopted; endpoint consolidation, tight aggregation of channels on a single board. This contribution describes the status and performances of the NaNet project, whose goal is the design of a family of FPGA-based PCIe network interface cards. The efforts of the team are focused on implementing a low-latency, real-time data transport mechanism between the board network multi-channel system and CPU and GPU accelerators memories on the host. Several opportunities concerning technical solutions and scientific applications have been explored: NaNet-1 with a single GbE I/O interface, and NaNet-10, offering four 10GbE ports, for activities related to the GPU-based real-time trigger of NA62 experiment at CERN; NaNet^3, with four 2.5 Gbit optical channels, developed for the KM3NeT-ITALIA underwater neutrino telescope.

1. – Introduction

Modern High Energy Physics experiments work at high interaction rates and an efficient trigger mechanism is often mandatory to pick potentially interesting events among crowded backgrounds. For the purposes of this discussion, two broad categories can be identified in the varied scenario of trigger systems: on-line and off-line.

The off-line triggers perform the analysis once data reaches the final destination, usually a PC-farm; in this case the primary obstacles to face are: i) database and storage management, to guarantee an efficient organization of recorded data, ii) high bandwidth

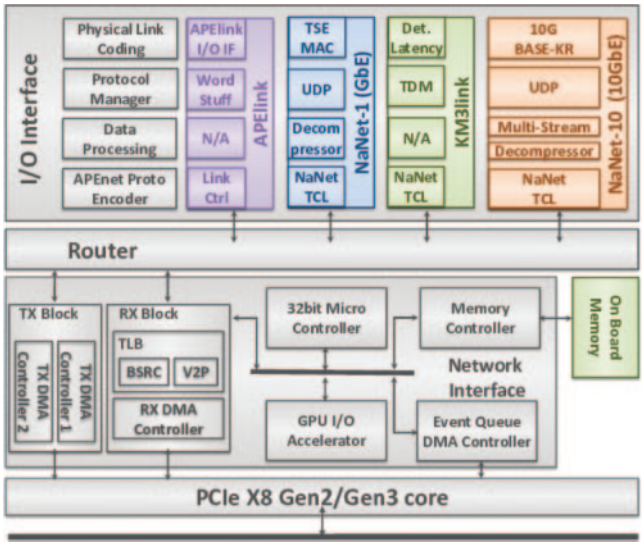

Fig. 1. – NaNet PCIe Network Interface Card family architecture.

(and low-latency as well) network adaptor, to handle large amounts of data in a fast way, iii) great computing power, to keep analysis time low and iv) endpoint consolidation capability, to keep to a minimum the number of final devices and reduce the operating cost of the server host infrastructure.

In the context of on-line analysis, the data flow is processed along the path; here the challenging parameters are: i) low and (almost) deterministic latency —full control of the time fluctuations in moving data towards the memory of the computing device— ii) high throughput of the trigger processor and iii) large computing power but with a deterministic execution behaviour. Available choices for data handling are: i) full hardware module —ASICs or FPGAs— implementing quite simple but fast and valuable trigger primitives, ii) exploitation of COTS hardware with trigger applications running on CPUs or many-core accelerators, such as GPUs.

A network interface component tasked with managing the data flow from the detector readout to the adopted computing device plays a key role in all the aforementioned cases.

2. – NaNet

The NaNet [1] project aims at designing and building a family of FPGA-based PCIe network interface cards (NIC) providing a bridge between the readout of the detectors and the computing nodes of a PC farm by means of a real-time data transport mechanism. The NaNet architecture is built upon the assurance of the complete control of the latency. This feature is achieved by implementing in hardware the data transfer tasks. In this way, latency fluctuations can be minimized — a core issue for an on-line trigger setup — until a completely deterministic behaviour for the communication management is reached, devoid of OS jitter effects. Consequently, hardware handling of the data path allows for a low-latency and high-bandwidth implementation of the data transport system.

In fig. 1 an overview of the NaNet design is sketched. The I/O interface is responsible for the network stack and is the distinctive trait of the NaNet board variant matched

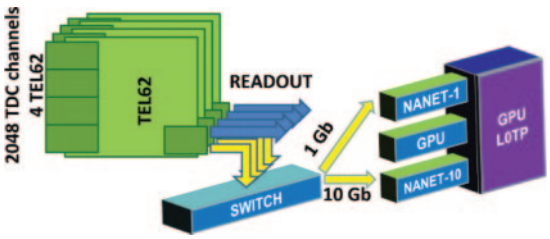

Fig. 2. – Pictorial view of GPU-based Trigger.

to the use case requirements. The Physical Link Coding and Protocol Manager IPs can be modified to match the detector readout interface of each experiment allowing for a reconfiguration of the channel. The former manages the line coding, alignment process, bit rate, transmission mode, media access and logical link control — *i.e.* 8B10B, 10/100/1000 MAC, Altera Deterministic Latency, Altera 10GBASE-KR PHY and the 10Gbps MAC [2] — while the latter is in charge of handling the communication protocol — *i.e.* standard UDP, Word-stuff and Time Division Multiplexing of proprietary APElink and KM3link channels. The Data Processing module can perform application-dependent modifications on the data stream on the fly — *e.g.* data reshuffling towards more device-friendly formats or (de)compression. The Transmission Control Logic (NaNet TCL) runs a protocol translation optimizing the PCIe transactions and generates the virtual addresses for appropriate interaction with the Linux Kernel. The router module consolidates the endpoints by multiplexing up to 10 data flows at 2.8 GB/s using a dimension-order routing algorithm. Finally, the Network Interface achieves Zero-copy networking, moving data to/from application memory by means of a DMA engine implementing the RDMA protocol for both CPU and GPU — this latter supports GPUDirect V2/RDMA by nVIDIA.

3. – NaNet at the NA62 CERN Experiment: data transport for the RICH Detector GPU-based L0 Trigger

Investigating the feasibility of a GPU-based L0 trigger system for the NA62 experiment (GPU_L0TP) led us to focus on the ring-shaped hit patterns reconstruction in the RICH detector of the NA62 experiment [3]. Here the detector identifies pions and muons with momentum in the range between 15 GeV/c and 35 GeV/c. Čerenkov light is reflected by a composite mirror with a focal length of 17 m and focused onto two separated spots equipped with $\sim$ 1000 photomultipliers (PM) each.

Data from the PMs are gathered by four readout boards (TEL62, Trigger Electronics for NA62), each sending primitives to the GPU_L0TP (fig. 2) by GbE UDP streams.

Communication main requirement is the deterministic response latency of GPU_L0TP; the available time budget forces both communication and computation to be under 1ms.

Refined primitives from the GPU-based computation are sent to the central L0 processor, where information from other detectors is taken into account in the trigger decision.

In 2015 the GPU-based trigger at CERN includes 2 TEL62 boards connected to a HP2920 switch and a NaNet-1 board with a TTC HSMC daughtercard plugged into a server made of a X9DRG-QF dual socket motherboard populated with Intel Xeon E5-2620 @2.00 GHz CPUs (*i.e.* Ivy Bridge architecture), 32 GB of DDR3 RAM and a Kepler-class nVIDIA K20c GPU.

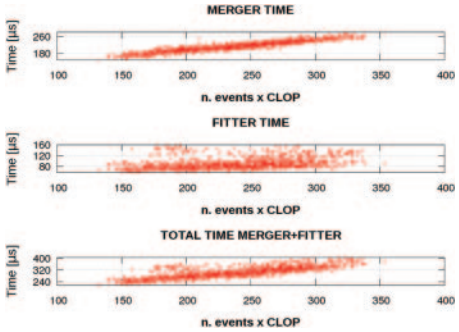

Fig. 3. – Multi-ring reconstruction.

Such a system allows testing of the whole chain: the data events move towards the GPU-based trigger through NaNet-1 by means of the GPUDirect RDMA interface. All data landing within a configurable time frame are gathered and organized in a parametrizable Circular List Of Persistent buffers (CLOP) in GPU memory. This time frame must be always shorter or equal to how long multi-ring reconstruction takes on the GPU, to be sure that buffers are not overwritten before they are consumed if one wants to avoid memory copies. Events are timestamped; those sharing a time-window but coming from different boards are fused into one event describing the PMs status in the RICH detector.

The multi-ring reconstruction GPU kernel is based on the histogram algorithm: the XY-plane is divided into a grid and the distances from gridpoints to hits of the physics event are put into bins so that those whose contents exceed a threshold value identify rings. The computing kernel implements the histogram fitter with a single step (*i.e.* over an 8×8 grid only) and is executed on a full CLOP buffer as soon as the NIC signals to the host application that the buffer is available to be consumed. Data selection might benefit from taking into account the parameters of Čerenkov rings, whose values are determined using the coordinates of activated PMs.

Results are in fig. 3; the CLOP sizes lie on the X-axis measured as number of received events while on the Y-axis the latencies of different stages are shown. Events coming from 2 readout boards, for a gathering time of $400\,\mu$s, and parameters like events rate (collected with a beam intensity of 4×10^{11} protons per spill), a CLOP's size of 8 KB, time frame was chosen so that we could test the online behaviour of the trigger chain. Since the merge operation requires synchronization and serialization and does not exhibit much parallelism, as such it is an ill-suited problem to the GPU architecture. Again, high latency of this task suggests offloading its duty to a hardware implementation.

4. – NaNet at KM3NeT-IT

KM3NeT-IT [4] is an underwater experimental apparatus for detection of high energy neutrinos in the TeV–PeV range. It consists of an array of photomultipliers (PMTs) exploiting the Čerenkov effect produced by charged particles propagating in sea water.

The final assessment of the experiment foresees the installation of 8 detection units called *tower* [5], each one consisting of 14 floors vertically spaced 20 m apart and a baseline. Each floor is 8 m long and hosts 6 Optical Modules (OM) —each containing a ten-inch photomultiplier (PMT) and the front-end module (FEM) to acquire and transmit the pulse generated by the PMT—, 2 hydrophones to reconstruct in real-time

the OM position and a board, called *Floor Control Module* (FCM). The FCM collects OM-produced data and manages the communication between the on-shore laboratory and the underwater devices through an optical fiber.

In order to reconstruct the particle tracks, accurate knowledge of spatial and temporal distribution of the hits is mandatory. This implies a continuous tracking of OMs position and a system-wide distribution of a common timing. As regards the last constraint, the endpoint board in the laboratory is in charge of distributing a common clock all over the system, and data frames labelled with a "time stamp" are encapsulated in a synchronous link protocol which embeds clock and data with deterministic latency.

The customization for the KM3NeT-IT experiment of NaNet board is called NaNet^3 [6]; by means of a real-time data transport mechanism it is able to manage multiple FCM data channels, allowing for more cost-effective and efficient scaling infrastructure. NaNet^3 is implemented on the Terasic DE5-net board, equipped with an Altera Stratix V FPGA, four I/O channels with maximum capability of 800 Mbps each and a PCIe Gen2 $\times$ 8 edge connector. The card implements a synchronous link protocol with deterministic latency at the physical level and a Time Division Multiplexing (TDM) protocol at the data level (see fig. 1). A GPS clock acts as reference and is used for the optical link transmission from the on-shore board towards the underwater FCM. GPS signals —*i.e.* clock and IRIG data— are received from the two SMA connectors on board. Incoming TDM data —*i.e.* hydrophones and photomultipliers data— are translated in a packet-based protocol, properly handled —*i.e.* computing the destination virtual address— and then sent to the computing node memory through a PCIe DMA write process. The outbound flow consists instead of slow-control data for the underwater apparatus; data addressed to the off-shore devices are stored in registers (one for each underwater device) through PCIe target mode and a custom TX module is in charge of dispatching the messages to the proper channel. One data frame is transmitted every 125 μs, with a GPS tagging every frame with a 12.5 ns precision timestamp.

5. – Conclusion

Our envisioned design of FPGA-based NICs with hardware-controlled latency and stream processing capabilities proved effective in two TRIDAQ contexts. Besides the high performance as a PCIe NIC, the key enabling features of NaNet provided acceleration both at communication protocol management and data stream processing stages, resulting in an integrated Network/Processing low-latency accelerator suitable for real-time (GPU) processing systems. The FPGA technology scaling endorses our current effort in further developing the design towards larger aggregated bandwidth and processing capabilities, to face the evergrowing requirements imposed by HEP experiments.

REFERENCES

[1] Ammendola R. *et al.*, *J. Instrum.*, **9** (2014) C02023.
[2] "Altera Transceiver PHY IP Core User Guide." `https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/ug/xcvr_user_guide.pdf`.
[3] Lamanna G., *J. Phys. Conf. Ser.*, **335** (2011) 012071.
[4] Margiotta A., *J. Instrum.*, **9** (2014) C04020.
[5] Nicolau C. A. *et al.*, *EPJ Web of Conferences*, **116** (2016) 05011.
[6] Ammendola R. *et al.*, *EPJ Web of Conferences*, **116** (2016) 05008.